Method for Screening Peptide Fragment Ion Mass Spectra Prior to Database Searching

Roger E. Moore, Mary K. Young, and Terry D. Lee
Beckman Research Institute of the City of Hope, Duarte, California, USA

A methodology is described for screening fragment ion spectra of peptides prior to database searching for protein identification. A software routine written in the Perl programming language was used to analyze data from previous Sequest database searches and develop a set of statistical descriptors that could be used to identify spectra not likely to yield useful results in a database search. A second Perl program used an evolutionary algorithm to optimize the criteria for each statistical descriptor and generate a formula for determining spectral quality. This formula was used by a third Perl program to screen data sets from four independent liquid chromatography tandem mass spectrometry runs. On the average, use of the screening program reduced the time required for a database search by 1/2 with little loss of useful information from the database search results. (J Am Soc Mass Spectrom 2000, 11, 422–426) © 2000 American Society for Mass Spectrometry

Identification of proteins by enzyme digestion, mass spectral analysis, and database searching has become a standard and very powerful technique. Three main approaches to database searching have been developed based on peptide mass fingerprinting [1–3], peptide mass plus partial sequence [4], and full fragmentation pattern matching [5]. Methods based on full fragmentation pattern matching are potentially the most powerful because they take advantage of all available information, but require much more computer processor time.

Using Sequest, a commercially available searching program based on full fragment pattern matching, to search a nonredundant database is comparatively slow. Sequest takes approximately 50 s to search one tandem mass spectrometry (MS/MS) spectrum against the OWL database (312,942 proteins) using a single 450 MHz Pentium II Xeon processor. A mass spectrometer with an on-line separation and data-dependent analysis can generate data 10 to 50 times faster. Much of the data from such an analysis are of very low quality as the mass spectrometer can continue to generate data even when no peptides are being analyzed. To reduce this problem, attempts have been made to automatically screen spectra to eliminate those derived from known contaminants [6]. However, this does nothing to eliminate spectra derived from background noise, or peptides that do not fragment well. It is possible to manually sort through the data to eliminate spectra that are unlikely to generate positive database search results. Using trained human analysts to save computer processing time is not a worthwhile solution, but suggests that there are quantifiable differences between promising and unpromising spectra. In this report, we describe a series of simple software routines that have been used to analyze the results of old searches to identify these quantifiable differences. Based on the results of those studies, a computer program was written to screen out unpromising spectra from a set submitted for database searching.

Experimental Methods

Data for the study were collected from old Sequest (Finnigan MAT, San Jose, CA) search files generated by the Mass Spectrometry Core Facility, Beckman Research Institute, City of Hope. The mass spectral data were generated using a Finnigan LCQ ion trap mass spectrometer, and were acquired in the centroid mode. The data extraction and screening programs were written in Active Perl for Windows v. 5.005, available from the Comprehensive Perl Archive Network (http://www.cpan.org). Extracted data was further manipulated using Microsoft (Redmond, Washington) Excel97. All Perl programs described here are available from the authors (http://www.cityofhope.org/immunology/download.html).

The Perl script *fullstats.pl* was written to extract descriptive statistics from MS/MS spectra submitted to Sequest and the cross correlation score from the corresponding search results. The statistics extracted were the total number of peaks, base peak intensity, total ion current (TIC), standard deviation of peak intensity, and the fraction of ions exceeding fixed relative abundances. The descriptive statistics and cross-correlation score were output to a delimited text file that was then

Address reprint requests to Terry D. Lee, Division of Immunology, Beckman Research Institute of the City of Hope, 1450 E. Duarte Rd., Duarte, CA 91010. E-mail: tdlee@coh.org

1044-0305/00/$20.00
PII S1044-0305(00)00097-0

Received September 14, 1999
Revised December 13, 1999
Accepted December 13, 1999

imported into Excel for further manipulation. Two additional Perl scripts, *cutoff.pl* and *weighted.pl*, used an evolutionary algorithm approach to generate selection criteria from the statistical data. The selection criteria were then incorporated into the Perl script *winnow.pl*, which creates a list of accepted files, a list of rejected files, and a log file describing why each file was accepted or rejected.

Results and Discussion

Developing selection criteria based on the results of old searches has several advantages. The data are readily available and represent a tremendous investment of resources. More importantly, data from previous searches are an accurate representation of the performance of the complete system that generated it. A number of factors, including the type and model of mass spectrometer, system parameters such as collision energy, and the specific search parameters used can have an impact on the search results. An analysis based on old data inherently incorporates all this information as well as certain limitations to the database searching approach. This means that the screening parameters generated will be optimized for the exact system in use, but may make the parameters less useful for screening spectra generated under substantially different conditions. For the work described here, results from searches over 10,000 individual MS/MS spectra were arbitrarily divided into two data sets of roughly equal size. Selection criteria developed by analyzing one data set were then tested using the other data set.

Sequest cannot produce a positive match for a spectrum of a peptide that is either not in the database being searched or contains unanticipated posttranslational modifications. Any Sequest search may contain false negative matches because the peptide sequence is not in the database being searched. The corresponding problem of false positive matches can be largely eliminated by choosing an appropriate cross-correlation score cutoff.

To avoid problems caused by false negative matches, a two-step approach was chosen to develop selection criteria. Spectra were separated into "good" and "bad" groups based on the cross-correlation score from the Sequest search. From our own experience and work published by others [7], the dividing line used was a score of 2.0. The optimum set of selection criteria was then defined to be one that selected a specified percentage of good spectra while including as few bad spectra as possible. Because the possible space of selection criteria was too large for an exhaustive search, we chose to use an evolutionary algorithm approach. An initial standard set of selection criteria was generated, and the number of bad spectra it selected was determined. The various criteria were then randomly modified, and the resulting number of bad spectra included by the modified criteria was compared to the number selected by the standard set. If the new criteria selected fewer bad spectra, they became the new standard set. The process was then repeated 1000 times to generate an optimized set of criteria. After a certain amount of experimentation it was found that 1000 cycles were sufficient to produce optimum results. The process of generating optimized criteria was repeated 20 times for each set of data and the best resulting criteria were selected. It was necessary to repeat the process of generating optimized criteria to ensure that the criteria generated were close to the global optimum of the data space and not a local optimum. Producing criteria by focusing on the common features of good spectra tends to minimize the potential problems caused by the high level of false negative matches. Any set of criteria that includes a high percentage of positive matches will also tend to include a high percentage of false negative matches, as their spectral characteristics are inherently similar.

The statistics extracted from the input MS/MS spectra were chosen to include the factors that human analysts use in judging whether a spectrum is likely to produce a positive match. These factors include the intensity of the spectrum, the total amount of data present in the spectrum, and factors relating to the distribution of peak intensities. The criteria chosen were total ion current and base peak intensity to represent spectral intensity, the number of peaks as a measure of the amount of data in the spectrum, and the standard deviation of peak intensities and the fraction of peaks exceeding fixed relative abundances to represent the distribution of peak intensities. Some of the criteria chosen, particularly the base peak intensity, TIC, and number of peaks, varied over several orders of magnitude in value. It was found that these factors did not work well when directly incorporated into a weighted formula, either dominating all other criteria or not being incorporated into the formula at all. Using the natural logarithm of these factors eliminated this problem. It was also found that a comparatively large data set was needed to get high quality results. In a data set containing only 1000 total spectra and only about 200 good spectra, a handful of anomalous spectra can severely bias the final results. At least 5000 spectra seemed to be a practical minimum for generating new formulas, meaning that 10,000 spectra were needed to have equally sized data sets for generating and testing new formulas. No attempt was made to incorporate measures relating to the distribution of peaks on the mass axis.

Two different methods of utilizing the selection criteria were used, which were designated as the cutoff method and the weighted method. In the cutoff method, each descriptive statistic had its own cutoff value. The cutoffs could be either low cutoffs, in which spectra had to exceed the cutoff value to be selected, or high cutoffs, in which spectra had to fall below the cutoff value. Spectra had to meet all cutoff values to be selected. In the weighted method, the values for various statistics were combined in a linear formula to produce

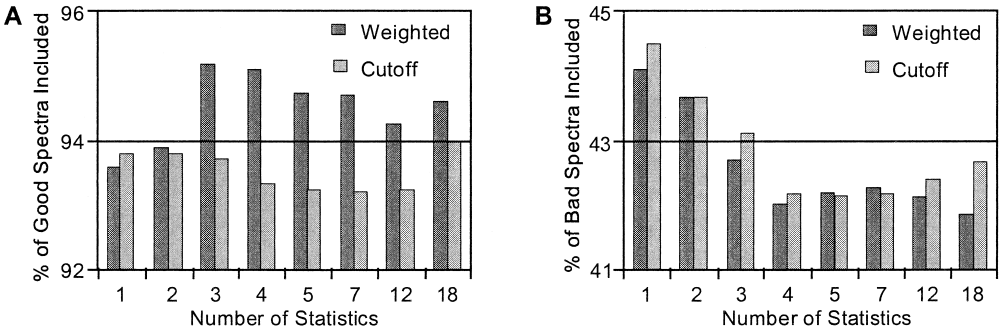

Figure 1. Percentage of (**A**) good spectra and (**B**) bad spectra included as a function of number of statistics used to develop the selection criteria. The ln(number of peaks), [ln(P)] was used for 1 statistic. For 2 statistics, ln(P) and ln(total ion current) [ln(T)] were used. For 3 statistics, ln(P), ln(T), and fraction of peaks >1% (F_1) were used. For 4 statistics, ln(P), ln(T), F_1, and fraction of peaks >20% (F_{20}) were used. For 5 statistics, ln(P), ln(T), F_1, F_{20}, and the fraction of peaks >50% (F_{50}) were used. For 7 statistics, the fractions of peaks >7% and >40% were added to the 5 statistics. For 12 statistics, the fraction of peaks >3%, 13%, and 30%, ln(base peak intensity), and the ratio of standard deviation to average peak intensity were added to the 7 statistics. For 18 peaks, the fraction of peaks >2%, 5%, 10%, 17%, 25%, and 35% were added to the 12 statistics.

a single composite score with a single cutoff value for selection.

Each method was tested using different numbers of descriptive statistics, and the selection criteria developed were applied to an independent data set. The weighted approach was somewhat superior to the cutoff approach, selecting a higher percentage of good spectra and a lower percentage of bad spectra when using the same set of statistics (Figure 1). The single most important statistic was found to be the natural logarithm of the number of peaks in the spectrum. Adding three additional statistics, the natural logarithm of total ion current and the fraction of peaks exceeding 1% and 20% relative abundance, improved the results noticeably. Including any of the other statistics that were considered did not improve the results. Because the results using the weighted method were superior to those using the cutoff method, all further experiments were performed using the weighted method only.

The weighted approach was then tested by varying the specified percentage of good spectra to be selected. The selection criteria developed using one data set were then applied to the second data set to test their practical applicability. The percentage of bad spectra selected dropped significantly as a few percent of the good spectra were not selected (Figure 2). When the percentage of good spectra selected was set below about 95%, the decrease in the number of bad spectra selected was less dramatic. As a result of this data, it was determined that a good choice of formula for screening good from bad spectra would be

$$0.25728*\ln(P) + 0.11836*\ln(T) + 0.39309*F_1 - 0.23127*F_{20} \geq 2.98546$$

where P is the number of peaks in the spectrum, T is the total ion current, F_1 is the fraction of peaks exceeding 1% relative abundance, and F_{20} is the fraction of peaks exceeding 20% relative abundance.

This formula was incorporated into the Perl program *winnow.pl*, which is used to analyze all the MS/MS spectra in an LC/MS run and select those to be used for Sequest searching. The value of the *winnow.pl* program was tested by running Sequest searches on four new, independent data sets with and without *winnow.pl* screening. The time saved for each run was roughly proportional to the ratio of bad spectra to the total number of spectra (Table 1). Results were similar for peptide mixtures generated using either trypsin or Asp N as the proteolytic enzyme. It is also important to note that the screening program works just as well for data sets with only a few good spectra as it does for sets containing many good spectra. For the combined data sets, the use of *winnow.pl* cut the time spent searching by 1/3 with the loss of only one good spectrum out of 107. Only a few seconds are needed for *winnow.pl* to screen each data set, which is insignificant compared to

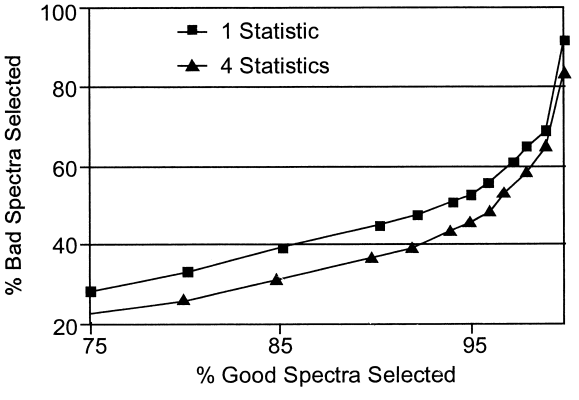

Figure 2. Percentage of bad spectra selected as a function of the percentage of good spectra selected using 1 statistic [ln(P)], and 4 statistics [ln(P), ln(T), F_1, and F_{20}].

Table 1. The effect of using *winnow.pl* to screen spectra from four different LC/MS runs on peptide mixtures obtained using trypsin and Asp N endoproteases. Two different versions of *winnow.pl* used either a single formula for all spectra or multiple formulae depending on parent ion charge state. All times are in minutes

				Search time						
					With screening		Time saved		Good spectra lost	
Sample	Enzyme	Total spectra	Good spectra	No screening	Single	Multiple	Single	Multiple	Single	Multiple
1	Trypsin	304	56	267	219	165	48 (18%)	102 (38%)	0	3
2	Trypsin	248	34	232	162	121	70 (30%)	111 (48%)	0	2
3	Asp N	179	9	132	54	24	78 (59%)	108 (82%)	1	2
4	Asp N	157	8	111	62	46	49 (44%)	65 (59%)	0	1
Total		888	107	742	497	356	245 (33%)	386 (52%)	1	8

the time required for the Sequest search. In addition to the list of spectra selected for database searching, *winnow.pl* also makes a list of the rejected spectra that can be searched later if desired. Thus, using *winnow.pl* does not increase the time to do a search even if the decision is made later to analyze all of the spectra. In the few instances where this has been done, no additional information on the sample was obtained by expanding the search to include the rejected files.

A further refinement of the technique incorporated information about the molecular weight and charge state of the parent ion in addition to the four criteria mentioned above. The information about charge state was incorporated into the formula in several different ways. In one approach, spectra from parent ions with different charge states were separated and different criteria were developed for each parent ion charge state. Another took essentially the same approach but grouped spectra from +2 and +3 parent ions, which seem to behave similarly for Sequest searching. A third approach grouped all spectra but included the parent ion charge state as an explicit criterion. The decision to split the data by charge state necessitated using a larger data set consisting of about 20,000 individual spectra and again divided arbitrarily into equally sized training and testing data sets. MS/MS spectra derived from parent ions with charge greater than 3 were excluded from the study, as they were extremely rare in the data set used and rarely gave positive search results. Each data set was tested both with and without the natural logarithm of the parent ion mass as a criterion. As shown in Figure 3, separating MS/MS spectra from singly charged parent ions from MS/MS spectra from doubly and triply charged parent ions significantly improved the quality of screening, but only if the parent ion mass was incorporated as a criterion. Further separating spectra derived from doubly and triply charged parent ions did not improve the results. A further analysis of the data indicates that the parent ion mass was a significant criterion only for MS/MS spectra derived from singly charged ions (data not shown). Results for MS/MS spectra derived from doubly and

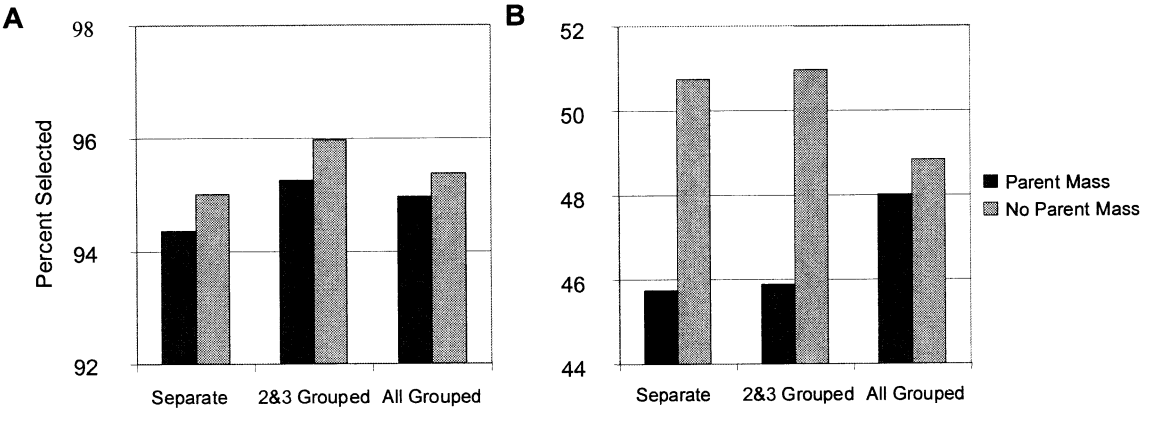

Figure 3. Percentage of (**A**) good spectra and (**B**) bad spectra included using different information about parent ion mass and charge state. Information about the charge state was incorporated by completely separating spectra by parent ion charge state (separate), separating spectra derived from singly charged parent ions from those derived from doubly and triply charged spectra (2 & 3 grouped), or grouping all spectra but including parent ion charge state as an explicit statistic (all grouped). Each data set was examined both with and without the natural log of parent ion mass as a statistic. Incorporating parent ion mass and separating out spectra derived from singly charged parent ions significantly improved the results by excluding more bad spectra. Treating spectra from doubly and triply charged parent ions separately had little effect.

triply charged parent ions were actually slightly better when not incorporating parent mass data, but the differences were not significant. The optimized formulas for acceptance were

$$z = 1 \qquad 0.4083*\ln(M) + 0.2013*\ln(P) + 0.0490*\ln(T) + 0.1513*F_1 - 0.1901*F_{20} \geq 4.5251$$

$$z = 2, 3 \qquad 0.1242*\ln(P) + 0.1437*\ln(T) + 0.4078*F_1 - 0.3243*F_{20} \geq 2.7037$$

$$z > 3 \qquad \text{not accepted}$$

where z is the parent ion charge state, M is the parent ion molecular weight, P is the number of peaks in the spectrum, T is the total ion current, and F_1 and F_{20} are the fraction of peaks exceeding 1% and 20% relative abundance, respectively.

These formulas were incorporated into the screening program *winnow.pl* and the new version was tested by analyzing the same four data sets as were analyzed using the first version. As shown in Table 1, the version of *winnow.pl* incorporating multiple formulas cut search times even more than the version using a single formula, 1/2 of total time instead of 1/3. The decreased search time did come at the cost of a small decrease in the number of good spectra. The multiple formulas excluded 8 of 107 good spectra instead of the 1 good spectrum lost using the single formula.

Conclusions

Prescreening of spectra before database searching is an effective method of cutting computing time. It eliminates a substantial number of bad spectra while keeping almost all of the good ones. Because database identification of proteins is an inherently robust process, elimination of any one spectrum will not generally compromise the correct identification of a protein. The computational savings from not searching unpromising spectra can also be used to look for modified amino acids in the spectra that are searched, potentially improving the quality of the overall results. The approach to screening spectra presented here is also complementary to approaches that search for spectra matching known contaminants.

This methodology is generally applicable. Differences between mass spectrometers or the nature of the information required to solve the problem may make it desirable to change the selection criteria which is readily done using these programs. As long as there is a similar way of dividing spectra into good and bad, the same approach could be adapted to other database searching techniques. The approach may also be usable to determine other significant parameters. By varying the cross correlation value used as a dividing line between good and bad spectra, it might be possible to find an objective standard for the cutoff.

Acknowledgments

This work was supported in part by grants from the Public Health Services (NIH RR06217 and CA33752).

References

1. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. USA* **1993,** *90*, 5011–5015.
2. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993,** *195*, 58–64.
3. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993,** *3*, 327–332.
4. Mann, M.; Wilm, M. *Anal. Chem.* **1994,** *66*, 4390–4399.
5. Eng, J. K.; McCormack, A. L.; Yates, J. R. III *J. Amer. Soc. Mass Spectrom.* **1994,** *5*, 976–989.
6. Yates, J. R. III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998,** *70*, 3557–3565.
7. Ducret, A.; Van Oostveen, I.; Eng, J. K.; Yates, J. R. III; Aebersold, R. *Protein Sci.* **1998,** *7*, 706–719.